

DATA REPPLICATION PROTOCOL

INVENTORS

Dean Bernard Jacobs
Reto Kramer
Ananthan Bala Srinivasan

CERTIFICATE OF MAILING BY "EXPRESS MAIL" UNDER 37 C.F.R. § 1.10

"Express Mail" mailing label number: EL 622 697 783 US
Date of Mailing: October 11, 2001

I hereby certify that this correspondence is being deposited with the United States Postal Service, utilizing the "Express Mail Post Office to Addressee" service addressed to **Box PATENT APPLICATION, Commissioner for Patents, Washington, D.C. 20231** and mailed on the above Date of Mailing with the above "Express Mail" mailing label number.


Matthew A. Mahling
Signature Date: October 11, 2001

DATA REPLICATION PROTOCOL

INVENTORS:

Dean Bernard Jacobs
Reto Kramer
Ananthan Bala Srinivasan

CLAIM OF PRIORITY

[0001] This application claims priority to U.S. Provisional patent application No. 60/305,986, filed July 16, 2001, entitled DATA REPLICATION PROTOCOL, incorporated herein by reference.

CROSS-REFERENCE TO RELATED APPLICATION

[0002] The following application is cross-referenced and incorporated herein by reference:

[0003] U.S. Patent Application No. _____ entitled "LAYERED ARCHITECTURE FOR DATA REPLICATION," inventors Dean Bernard Jacobs, Reto Kramer, and Ananthan Bala Srinivasan, filed October 11, 2001.

COPYRIGHT NOTICE

[0004] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent

document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

TECHNICAL FIELD

[0005] The invention relates generally to a system for transferring data. The invention relates more specifically to a system and method for replicating data over a network.

5

BACKGROUND

[0006] There are several types of distributed processing systems. Generally, a distributed processing system includes a plurality of processing devices, such as two computers coupled through a communication medium. One type of distributed processing system is a client/server network. A client/server network includes at least two processing devices, typically a central server and a client. Additional clients may be coupled to the central server, there may be multiple servers, or the network may include only servers coupled through the communication medium.

[0007] In such a network environment, it is often desirable to send applications or information from the central server to a number of workstations and/or other servers. Often, this may involve separate installations on each workstation, or may involve separately pushing a new

library of information from the central server to each individual workstation and/or server. These approaches can be time consuming and are an inefficient use of resources. The separate installation of applications on each workstation or server also introduces additional potential sources of
5 error.

[0008] Ideally, the sending of information should be both reliable in the face of failures and scalable, so that the process makes efficient use of the network. Conventional solutions generally fail to achieve one or both of these goals. One simple approach is to have a master server
10 individually contact each slave and transfer the data over a point-to-point link, such as a TCP/IP connection. This approach leads to inconsistent copies of the data if one or more slaves are temporarily unreachable, or if the slaves encounter an error in processing the update. At the other extreme are complex distributed agreement protocols, which require
15 considerable cross-talk among the slaves to ensure that all copies of the data are consistent.

BRIEF SUMMARY

[0009] The present invention includes a method for replicating data
20 from a master server to at least one slave or managed server, such as may be accomplished on a network. In the method, it may be determined whether the replication should be accomplished in a one or two phase method. If the replication is to be accomplished in a one phase method, a

version number may be sent that corresponds to the current state of the data on the master server. This version number may be sent to every slave server on the network, or only a subset of slave servers. The slave servers receiving the version number may then request that a delta be sent from
5 the master. The delta may contain data necessary to update the data on that slave to correspond to the current version number.

[0010] If the replication is to be accomplished in a two phase method, a packet of information may be sent from the master to each slave, or a subset of slaves. Those slaves may then respond to the master
10 server whether they can commit the packet of information. If at least some of the slaves can commit the data, the master may signal to those slave that they should process the commit. After processing the commit, those slaves may update to the current version number. If any of the slaves are unable to process the commit, the commit may be aborted.

15

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 is a diagram of a domain structure in accordance with one embodiment of the present invention.

20 [0012] Figure 2 is a diagram of layered architecture in accordance with one embodiment of the present invention.

[0013] Figure 3 is a diagram of a clustered domain structure in accordance with one embodiment of the present invention.

[0014] Figure 4 is a diagram of one phase process for a layered

architecture in accordance with one embodiment of the present invention.

[0015] Figure 5 is a diagram of two phase process for a layered architecture in accordance with one embodiment of the present invention.

[0016] Figure 6 is a flowchart for a one phase process in accordance
5 with one embodiment of the present invention.

[0017] Figure 7 is a flowchart for a two phase process in accordance
with one embodiment of the present invention.

DETAILED DESCRIPTION

10 [0018] The present invention provides for the replication of data or other information, such as from a master server, or “administration” server (“Admin server”), to a collection of slave servers, or “managed” servers. This replication can occur over any appropriate network, such as a conventional local area network or ethernet. In one embodiment, a master
15 server owns the original record of all data on the network, to which any updates are to be applied. A copy of the data, together with updates as they occur, can be transmitted to each slave server. One example application involves the distribution of configuration information from an Admin server to a collection of managed servers.

20 [0019] In one system in accordance with the present invention, it may be necessary for a service, such as a Data Replication Service (DRS), to distribute configuration and deployment information from an Admin Server to managed servers in the appropriate domain. Large data items

can be distributed over point-to-point connections, such as Transmission Control Protocol ("TCP"), since a multicast protocol like User Datagram Protocol ("UDP") does not have flow control, and can overwhelm the system. Remote Method Invocation (RMI), Hypertext Transfer Protocol
5 (HTTP), or a similar protocol may be used for point-to-point connections.

[0020] Managed servers can also persistently cache data on local disks. Without such caching, an unacceptable amount of time may be required to transfer the necessary data. The ability of the managed servers to cache is important, as it increases the speed of startup by reducing the
10 amount of startup data to be transferred. Caching can also allow startup and/or restart if the Admin Server is unreachable. Restart may be a more attractive option, and it may be the case that the Admin server directs a server to start. Caching, however, can provide the ability to start the domain without the Admin Server being available.

15 [0021] As shown in the domain structure 100 of **Figure 1**, an Admin Server 102 and at least one managed server 104 can comprise a domain 106. This domain 106 can be the administration unit for startup and shutdown. In one embodiment, a browser 108, or other user application or device, tells the Admin Server 102 to start. The Admin Server 102 then
20 tells all managed servers 104 in the domain 106 to start, and passes the appropriate configuration information. If a server goes down after the managed servers 104 have started, it may be desirable for that server to restart automatically, whether or not the Admin Server 102 is available.

Cached data can be useful for this purpose.

[0022] Updates to data on the Admin Server can be packaged as incremental deltas between versions. The deltas can contain configuration and/or other information to be changed. It may be preferable to update the configuration while the domain is running, as it may be undesirable to take the system offline. In one embodiment, the configuration changes happen dynamically, as they are pushed out by the Admin Server. Only the changes to the configuration are sent in the deltas, as it may be unnecessary, and unduly cumbersome, to send the full configuration each time.

[0023] A protocol in accordance with the present invention integrates two methods for the distribution of updates, although other appropriate methods may be used accordingly. These distribution methods may be referred to as a one-phase method and a two-phase method, and can provide a tradeoff between consistency and scalability. In a one-phase method, which may favor scalability, each slave can obtain and process updates at its own pace. Slaves can get updates from the master at different times, but can commit to the data as soon as it is received. A slave can encounter an error in processing an update, but in the one-phase method this does not prevent other slaves from processing the update.

[0024] In a two-phase method in accordance with the present invention, which may favor consistency, the distribution can be "atomic", in that either all or none of the slaves successfully process the data. There

can be separate phases, such as prepare and commit phases, which can allow for a possibility of abort. In the prepare phase, the master can determine whether each slave can take the update. If all slaves indicate that they can accept the update, the new data can be sent to the slaves to
5 be committed in the commit phase. If at least one of the slave servers cannot take the update, the update can be aborted and there may not be a commit. In this case, the managed servers can be informed that they should roll back the prepare and nothing is changed. Such a protocol in accordance with the present invention is reliable, as a slave that is
10 unreachable when an update is committed, in either method, eventually gets the update.

[0025] A system in accordance with the present invention can also ensure that a temporarily unavailable server eventually receives all updates. For example, a server may be temporarily isolated from the
15 network, then come back into the network without restarting. Since the server is not restarting, it normally will not check for updates. The server coming back into the network can be accounted for by having the server check periodically for new updates, or by having a master server check periodically to see whether the servers have received the updates.

20 **[0026]** In one embodiment, a master server regularly sends multicast "heartbeats" to the slave servers. Since a multicast approach can be unreliable, it is possible for a slave to miss arbitrary sequences of heartbeats. For instance, a slave server might be temporarily disconnected

from the network due to a network partitioning, or the slave server itself might be temporarily unavailable to the network due, causing a heartbeat to be missed. Heartbeats can therefore contain a window of information about recent updates. Such information about previous updates may be

5 used to reduce the amount of network traffic, as explained below.

[0027] There can be at least two layers within each master and each slave: a user layer and a system layer (or DRS layer). The user layer can correspond to the user of the data replication system. A DRS layer can correspond to the implementation of the data replication system itself. The

10 interaction of these participants and layers is shown in **Figure 2**.

[0028] As shown in the startup diagram **200** of **Figure 2**, the master user **202** and slave user **204** layers in this embodiment make downcalls into the master DRS **206** and slave DRS **208** layers, respectively. Such downcalls can, for example, take the form of:

15 *registerMaster(DID, verNum, listener)*
registerSlave(DID, verNum, listener)

where *DID* is an identifier taken from knowledge of well-known DIDs and refers to the object of interest, *verNum* is taken from the local persistent store as the user's current version number, and *listener* is an object that will handle upcalls from the DRS layer. The upcall can call a method on the listener object. The master can then begin to send heartbeats, or periodic deltas, with the current version number. A container layer **210** is shown, which can include containers adapted to take information from the slave

user **204**. Examples of possible containers include enterprise Java beans, web interfaces, and J2EE (Java 2 Platform, Enterprise Edition) applications. Other applications and/or components can plug into the container layer **210**, such as an administration client **212**. Examples of
5 update messaging between the User and DRS layers are shown for the one phase method in **Figure 4**, as well as for the two phase method in
Figure 5.

[0029] **Figure 4** shows one basic process **400** that may be used for a one-phase distribution approach in a layered architecture in accordance
10 with the present invention. In this process, the master user layer **402** makes a downcall **404** into the master DRS layer **406** to start a one phase distribution. This call can be to all slaves in the system, or only to a subset of slave servers. If the call is to a subset, the master user layer **402** can determine the scope of the update, or which slaves should receive the
15 update.

[0030] The master DRS layer begins multicasting heartbeats **408**, containing the current version number of the data on the master, to the slave DRS layer **410**. The slave DRS layer **410** requests the current version number **412** for the slave from the slave user layer **414**. The slave
20 user layer **414** then responds **416** to the slave DRS layer **416** with the slave version number. If the slave is in sync, or already is on the current version number, then no further requests may be made until the next update. If the slave is out-of-sync and the slave is in the scope of the update, the slave

DRS layer **410** can request a delta **420** from the master DRS layer **406** in order to update the slave to the current version number of the data on the master. The master DRS layer **406** requests **422** that the master user layer **402** create a delta to update the slave. The master user layer **402** then sends the delta **424** to the master DRS layer **406**, which forwards the delta **426** and the current version number of the master to the slave DRS layer **410**, which sends the delta **426** to the slave user to be committed. The current version number is sent with the delta in case the master has updated since the heartbeat **408** was received by the slave.

10 [0031] The master DRS layer **406** can continue to periodically send a multicast heartbeat containing the version number **408** to the slave server(s). This allows any slave that was unavailable, or unable to receive and process a delta, to determine that it is not on the current version of the data and request a delta **420** at a later time, such as when the slave comes 15 back into the system.

20 [0032] Figure 5 shows one basic process **500** that may be used for a two phase distribution approach in a layered architecture in accordance with the present invention. In this process, the master user layer **504** makes a downcall **504** into the master DRS layer **506** to start a two phase distribution. The master user layer **502** may again need to determine the scope of the update, and may set a "timeout" value for the update process.

[0033] The master DRS layer **506** sends the new delta **508** to the slave DRS layer **510**. The slave DRS layer **510** sends a prepare request

512 to the slave user layer 514 for the new delta. The slave user layer 514 then responds 516 to the slave DRS layer 510 whether or not the slave can process the new delta. The slave DRS layer forwards the response 518 to the master DRS layer 506. If the slave cannot process the request 5 because it is out-of-sync, the master DRS layer 506 makes an upcall 520 to the master user layer 502 to create a delta that will bring the slave in sync to commit the delta. The master user layer 502 sends the syncing delta 522 to the master DRS layer, which forwards the syncing delta 524 to the slave DRS layer 510. If the slave is able to process the syncing 10 delta, the slave DRS layer 510 will send a sync response 526 to the master DRS layer 506 that the slave can now process the new delta. If the slave is not able to process the syncing delta, the slave DRS layer 510 will send the appropriate sync response 526 to the master DRS layer 506. The master DRS layer 506 then heartbeats a commit or abort message 528 to 15 the slave DRS layer 510, depending on whether or not the slave responded that it was able to process the new delta. If all slave were able to prepare the delta, for example, the master can heartbeat a commit signal. Otherwise, the master can heartbeat an abort signal. The heartbeats also contains the scope of the update, such that a slave knows whether or not 20 it should process the information contained in the heartbeat.

[0034] The slave DRS layer forwards this command 530 to the slave user layer 514, which then commits or aborts the update for the new delta. If the prepare phase was not completed within a timeout value set by the

master user layer **502**, the master DRS layer **506** can automatically heartbeat an abort **528** to all the slaves. This may occur, for example, when the master DRS layer **506** is unable to contact at least one of the slaves to determine whether that slave is able to process the commit. The 5 timeout value can be set such that the master DRS layer **506** will try to contact the slave for a specified period of time before aborting the update.

[0035] For an update in a one-phase method, these heartbeats can cause each slave to request a delta starting from the slave's current version of the data. Such a process is shown in the flowchart of **Figure 6**. In this 10 basic process **600**, which may or may not utilize a layered architecture in accordance with the present invention, a version number for the current data on the master server is sent from a master server to a slave server **602**. The slave server determines whether it has been updated to the current version number **604**. If the slave is not on the current version, it will 15 request that a delta be sent from the master server to update the slave server **606**. When the delta is sent to the slave server, the slave server will process the delta in order to update the slave data to the current version **608**. The slave server will then update its version number to the current version number **610**.

20 [0036] For an update in a two-phase method, the master can begin with a prepare phase in which it pro-actively sends each slave a delta from the immediately-previous version. Such a process is shown in the flowchart of **Figure 7**. In this basic process **700**, which may or may not utilize a

layered architecture in accordance with the present invention, a packet of information is sent from the master to a slave server or slave servers **702**. Each slave server receiving the packet determines whether it can process that packet and update to the current version **704**. Each slave server 5 receiving the packet responds to the master server, indicating whether the slave server can process the packet **706**. If all slaves (to which the delta is sent) acknowledge successful processing of the delta within some timeout period, the master may decide to commit the update. Otherwise, the master server may decide to abort the update. Once this decision is made, 10 the master server sends a message to the slave server(s) indicating whether the update should be committed or aborted **708**. If the decision is to commit, each server processes the commit **710**. Heartbeats may further be used to signal whether a commit or abort occurred, in case the command was missed by one of the slaves.

15 [0037] A slave can be configured to immediately start and/or restart using cached data, without first getting the current version number from the master. As mentioned above, one protocol in accordance with the present invention allows slaves to persistently cache data on local disks. This caching decreases the time needed for system startup, and improves scalability by reducing the amount of data needing to be transferred. The protocol can improve reliability by allowing slaves to startup and/or restart if the master is unreachable, and may further allow updates to be packaged 20 as incremental deltas between versions. If no cache data exists, the slave

can wait for the master or can pull the data itself. If the slave has the cache, it may still not want to start out of sync. Startup time may be decreased if the slave knows to wait.

[0038] The protocol can be bilateral, in that a master or slave can
5 take the initiative to transfer data, depending upon the circumstances. For example, a slave can pull a delta from the master during domain startup. When the slave determines it is on a different version than the delta is intended to update, the slave can request a delta from its current version to the current system version. A slave can also pull a delta during one-
10 phase distribution. Here, the system can read the heartbeat, determine that it has missed the update, and request the appropriate delta.

[0039] A slave can also pull a delta when needed to recover from exceptional circumstances. Exceptional circumstances can exist, for example, when components of the system are out of sync. When a slave
15 pulls a delta, the delta can be between arbitrary versions of the data. In other words, the delta can be between the current version of the slave and the current version of the system (or domain), no matter how many iterations apart those versions might be. In this embodiment, the availability of a heartbeat and the ability to receive deltas can provide
20 synchronization of the system.

[0040] In addition to the ability of a slave to pull a delta, a master can have the ability to push a delta to a slave during two-phase distribution. In one embodiment, these deltas are always between successive versions of

the data. This two-phase distribution method can minimize the likelihood of inconsistencies between participants. Slave users can process a prepare as far as possible without exposing the update to clients or making the update impossible to roll back. This can include such tasks as checking the servers for conflicts. If any of the slaves signals an error, such as by sending a “disk full” or “inconsistent configuration” message, the update can be uniformly rolled back.

5 [0041] It is still possible, however, that inconsistencies may arise. For instance, there may be errors in processing a commit, for reasons such as an inability to open a socket. Servers can also commit and expose the update at different times. Because the data cannot reach every managed server at exactly the same time, there can be some rippling effect. The use 10 of multicasting can provide for a small time window, in an attempt to minimize the rippling effect. In one embodiment, a prepared slave will abort as an inability to open a socket. Servers can also commit and expose the update at different times. Because the data cannot reach every managed server at exactly the same time, there can be some rippling effect. The use 15 of multicasting can provide for a small time window, in an attempt to minimize the rippling effect. In one embodiment, a prepared slave will abort if it misses a commit, whether it missed the signal, the master crashed, etc.

15 [0042] A best-effort approach to multicasting can cause a slave server to miss a commit signal. If a master crashes part way through the commit phase, there may be no logging or means for recovery. There may be no way for the master to tell the remaining slaves that they need to 20 commit. Upon abort some slaves may end up committing the data if the version is not properly rolled back. In one embodiment, the remaining slaves could get the update using one-phase distribution. This might happen, for example, when a managed server pulls a delta in response to

a heartbeat received from an Admin server. This approach may maintain system scalability, which might be lost if the system tied down distribution in order to avoid any commit or version errors.

- 5 [0043] Each data item managed by the system can be structured to have a unique, long-lived domain identifier (DID) that is well-known across the domain. A data item can be a large, complex object made up of many components, each relevant to some subset of the servers in the domain. Because these objects can be the units of consistency, it may be desirable to have a few large objects, rather than several tiny objects. As an
- 10 example, a single data item or object can represent all configuration information for a system, including code files such as a config.xml file or an applicaiton-EAR file. A given component in the data item can, for example, be relevant to an individual server as to the number of threads, can be relevant to a cluster as to the deployed services, or can be relevant to the
- 15 entire domain regarding security certificates. A delta between two versions can consist of new values for some or all of these components. For example, the components may include all enterprise Java beans deployed on members of the domain. A delta may include changes to only a subset of these Java beans.
- 20 [0044] The “scope” of a delta can refer to the set of all servers with a relevant component in the delta. An Admin server in accordance with the present invention may be able to interpret a configuration change in order to determine the scope of the delta. The DRS system on the master may

need to know the scope in order to send the data to the appropriate slaves.

It might be a waste of time and resources to send every configuration update to every server, when a master may only need to only touch a subset of servers in each update.

- 5 [0045] To control distribution, the master user can provide the scope of each update along with the delta between successive versions. A scope may be represented as a set of names, referring to servers and/or clusters, which may be taken from the same namespace within a domain. In one embodiment, the DRS uses a resolver module to map names to addresses.
- 10 A cluster name can map to the set of addresses of all servers in that cluster. These addresses can be relative, such as to a virtual machine. The resolver can determine whether there is an intervening firewall, and return either an “inside” or “outside” address, relating to whether the server is “inside the firewall” as is known and used in the art. An Admin server or
- 15 other server can initialize the corresponding resolver with configuration data.

- 20 [0046] Along with the unique, long-lived domain identifier (DID) for each managed data item, each version of a data item can also have a long-lived version number. Each version number can be unique to an update attempt, such that a server will not improperly update or fail to update due to confusion as to the proper version. Similarly, the version number for an aborted two-phase distribution may not be re-used. The master may be able to produce a delta between two arbitrary versions given just the

version numbers. If the master cannot produce such a delta, a complete copy of the data or application may be provided.

[0047] It may be desirable to keep the data replication service as generic as possible. A few assumptions may therefore be imposed upon
5 the users of the system. The system may rely on, for example, three primary assumptions:

- the system may include a way to increment a version number
- the system may persistently store the version number on the master as well as the slave
- 10 • the system may include a way to compare version numbers and determine equality

These assumptions may be provided by a user-level implementation of a DRS interface, such as an interface "VersionNumber." Such an interface may allow a user to provide a specific notion and implementation of the
15 version number abstraction, while ensuring that the system has access to the version number attributes. In Java, for example, a VersionNumber interface may be implemented as follows:

```
20 package weblogic.drs;
public interface VersionNumber extends Serializable {
    VersionNumber increment();
    void persist() throws Exception;
    boolean equals(VersionNumber anotherVN);
    boolean strictlyGreaterThan(VersionNumber anotherVN);
}
25
```

A simplistic implementation of this abstraction that a user could provide to the system would be a large, positive integer. The implementation may also ensure that the system can transmit delta information via the network

from the master to the slaves, referred to in the art as being "serializable."

[0048] If using the abstraction above, it may be useful to abstract from a notion of the detailed content of a delta at the user level. The system may require no knowledge of the delta information structure, and
5 in fact may not even be able to determine the structure. The implementation of the delta can also be serializable, ensuring that the system can transmit delta version information via the network from the master to the slaves.

[0049] It may be desirable to have the master persistently store the
10 copy of record for each data item, along with the appropriate DID and version number. Before beginning a two-phase distribution, the master can persistently store the proposed new version number to ensure that it is not reused, in the event the master fails. A slave can persistently store the latest copy of each relevant data item along with its DID and version
15 number. The slave can also be configured to do the necessary caching, such that the slave may have to get the data or protocol every time. This may not be desirable in all cases, but may be allowed in order to handle certain situations that may arise.

[0050] A system in accordance with the present invention may
20 further include concurrence restrictions. For instance, certain operations may not be permitted during a two-phase distribution of an update for a given DID over a given scope. Such operations may include a one- or two-phase update, such as a modification of the membership of the scope on

the same DID, over a scope with a non-empty intersection.

[0051] In at least one embodiment, the master DRS regularly multicasts heartbeats, or packets of information, to the slave DRS on each server in the domain. For each DID, a heartbeat may contain a window of 5 information about the most recent update(s), including each update version number, the scope of the delta with respect to the previous version, and whether the update was committed or aborted. Information about the current version may always be included. Information about older versions can also be used to minimize the amount of traffic back to the master, and 10 not for correctness or liveness.

[0052] With the inclusion of older version information in a delta, the slave can commit that portion of the update it was expecting upon the prepare, and ask for a new delta to handle more recent updates. Information about a given version can be included for at least some fixed, 15 configurable number of heartbeats, although rapid-fire updates may cause the window to increase to an unacceptable size. In another embodiment, information about an older version can be discarded once a master determines that all slaves have received the update.

[0053] Multicast heartbeats may have several properties to be taken 20 into consideration. These heartbeats can be asynchronous or "one-way". As a result, by the time a slave responds to a heartbeat, the master may have advanced to a new state. Further, not all slaves may respond at exactly the same time. As such, a master can assume that a slave has no

- knowledge of its state, and can include that which the delta is intended to update. These heartbeats can also be unreliable, as a slave may miss arbitrary sequences of heartbeats. This can again lead to the inclusion of older version information in the heartbeats. In one embodiment, heartbeats
- 5 are received by a slave in the order they were sent. For example, a slave may not commit version seven until it has committed version six. The server may wait until it receives six, or it may simply throw out six and commit seven. This ordering may eliminate the possibility for confusion that might be created by versions going backwards.
- 10 [0054] As mentioned above, the domains may also utilize clustering, as shown in Figure 3 (Properties of Multicast Heartbeats slide). The general network topology for this embodiment is a collection of multicast islands, connected to a hub island containing the master. Multicast traffic may be forwarded point-to-point outward from the hub. Small deltas that
- 15 may be distributed in the one-phase method may be directly transmitted over multicast. In all other cases, deltas may be transmitted over point-to-point links. A tree-structured, point-to-point forwarding scheme may be overlaid on the hub-and-spoke multicast structure to reduce the bottleneck at the master.
- 20 [0055] In the domain diagram 300 of **Figure 3**, one or more of the managed servers 302 can be grouped into a multicast island, also referred to as a cluster 304. An Admin server 306 for the domain 308 acts as the master of the hub island 312, and is the entry point to the domain, such as

through a browser **310**. The Admin server **306** contacts one of the managed servers in the cluster, referred to as the cluster master. The Admin server in this embodiment can multicast a delta or message to each cluster master, with each cluster master then forwarding that delta or message by multicast to the other managed servers in that cluster. The cluster masters may not own any configuration information, instead receiving the information from the Admin server. In the event that a cluster master goes offline or crashes, another managed server in the domain may take over as cluster master. In this event, a mechanism may be put in place to prevent the offline server from coming back into the cluster as a second cluster master. This may be handled by the cluster or system infrastructure.

[0056] There can also be more than one domain. In this case, there can be nested domains or “syndicates.” Information can be spread to the domain masters by touching each domain master directly, as each domain master can have the ability to push information to the other domain masters. It may, however, be undesirable to multicast to domain masters.

[0057] In one-phase distribution, a master user can make a downcall in order to trigger the distribution of an update. Such a downcall can take the form of:

startOnePhase(DID, newVerNum, scope)

where *DID* is the ID of the data item or object that was updated, *newVerNum* is the new version number of the object, and *scope* is the

scope to which the update applies. The master DRS may respond by advancing to the new version number, writing the new number to disk, and including the information in subsequent heartbeats.

[0058] When a slave DRS receives a heartbeat, it can determine whether it needs a pull by analyzing the window of information relating to recent updates of interest. If the slave's current version number is within the window and the slave is not in the scope of any of the subsequent committed updates, it can simply advance to the latest version number without pulling any data. This process can include the trivial case where the slave is up-to-date. Otherwise, the slave DRS may make a point-to-point call for a delta from the master DRS, or another similar request, which may take the form of:

createDelta(DID, curVerNum)

where *curVerNum* is the current number of the slave, which will be sent back to the domain master or cluster master. To handle this request, the master DRS may make an upcall, such as *createDelta(curVerNum)*. This upcall may be made through the appropriate listener in order to obtain the delta and the new version number, and return them to the slave DRS. The new version number should be included, as it may have changed since the slave last received the heartbeat. The delta may only be up to the most recently committed update. Any ongoing two-phase updates may be handled through a separate mechanism. The slave DRS may then make an upcall to the slave user, such as *commitOnePhase(newVerNum, delta)*

and then advance to the new version number.

[0059] In order to trigger a two-phase update distribution, the master user can make a downcall, such as *startTwoPhase(DID, oldVerNum, newVerNum, delta, scope, timeout)*, where *DID* is the ID of the data item
5 or object to be updated, *oldVerNum* is the previous version number, *newVerNum* is the new version number (one step from the previous version number), *delta* is the delta between the successive versions to be pushed, *scope* is the scope of the update, and *timeout* is the maximum time-to-live for the job. Because the “prepare” and “commit” are synchronous, it may
10 be desirable to set a specific time limit for the job. The previous version number may be included to that a server on a different version number will not take the delta.

[0060] The master DRS in one embodiment goes through all servers in the scope and makes a point-to-point call to each slave DRS, such as
15 *prepareTwoPhase(DID, oldVerNum, newVerNum, delta, timeout)*. The slave can then get the appropriate timeout value. Point-to-point protocol can be used where the delta is large, such as a delta that includes binary code. Smaller updates, which may for example include only minor configuration changes such as modifications of cache size, can be done
20 using the one-phase method. This approach can be used because it may be more important that big changes like application additions get to the servers in a consistent fashion. The master can alternatively go to cluster masters, if they exist, and have the cluster masters make the call. Having

the master proxy to the cluster masters can improve system scalability.

[0061] In one embodiment, each call to a slave or cluster master produces one of four responses, such as "Unreachable", "OutOfSync", "Nak", and "Ack", which are handled by the master DRS. If the response

5 is "Unreachable", the server in question cannot be reached and may be queued for retry. If the response is "OutOfSync", the server may be queued for retry. In the meantime, the server will attempt to sync itself by using a pull from the master, so that it may receive the delta upon retry. If the response is "NoAck", or no acknowledgment, the job is aborted. This
10 response may be given when the server cannot accept the job. If the response is "Ack", no action is taken.

[0062] In order to prepare the slaves, a master DRS can call a method such as *prepareTwoPhase*. Upon receiving a "prepare" request from the master DRS, the slave DRS can first check whether its current
15 version number equals the old version number to be updated. If not, the slave can return an "OutOfSync" response. The slave can then pull a delta from the master DRS as if it had just received a heartbeat. Eventually, the master DRS can retry the *prepareTwoPhase*. This approach may be more simple than having the master push the delta, but may require careful

20 configuration of the master. The configuring of the master may be needed, as waiting too long for a response can cause the job to timeout. Further, not waiting long enough can lead to additional requests getting an "OutOfSync" response. It may be preferable to trigger the retry upon

completion of the pull request from the slave.

- [0063] If the slave is in sync, the slave can make an upcall to the client layer on the slave side, as deep into the server as possible, such as *prepareTwoPhase(newVerNum, delta)*. The resulting “Ack” or “Nak” that is returned can then be sent to the master DRS. If the response was an “Ack”, the slave can go into a special prepared state. If the response was a “Nak”, the slave can flush any record of the update. If it were to be later committed for some reason, the slave can obtain it as a one-phase distribution, which may then fail.
- 5
- [0064] If the master DRS manages to collect an “Ack” from every server within the timeout period, it can make a commit upcall, such as *twoPhaseSucceeded(newVerNum)*, and advance to the new version number. If the master DRS receives a “Nak” from any server, or if the timeout period expires, the master DRS can make an abort upcall, such as
- 10
- 15 *twoPhaseFailed(newVerNum, reason)*, and leave the version number unchanged. Here, *reason* is an exception, containing a roll-up of any “Nak” responses. In both cases, the abort/commit information can be included in subsequent heartbeats.
- [0065] At any time, the master DRS can make a cancel downcall,
- 20 such as *cancelTwoPhase(newVerNum)*. The master DRS can then handle this call by throwing an exception, if the job is not in progress, or acting as if an abort is to occur.
- [0066] If a prepared slave DRS gets a heartbeat indicating the new

version was committed, the slave DRS can make an upcall, such as *commitTwoPhase(newVerNum)*, and advance to the new version number.

If a prepared slave DRS instead gets a heartbeat indicating the new version was aborted, the slave can abort the job. The slave can also abort the job

5 when the slave gets a heartbeat where the window has advanced beyond the new version, the slave gets a new *prepareTwoPhase* call on the same data item, or the slave times out the job. In such a case, the slave can make an upcall, such as *abortTwoPhase(newVerNum)*, and leave the version number unchanged. This is one way to ensure the proper handling
10 of situations such as where a master server fails after the slaves were prepared but before the slaves commit.

[0067] The foregoing description of preferred embodiments of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to

15 the precise forms disclosed. Obviously, many modifications and variations will be apparent to the practitioner skilled in the art. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various
20 modifications that are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalence.